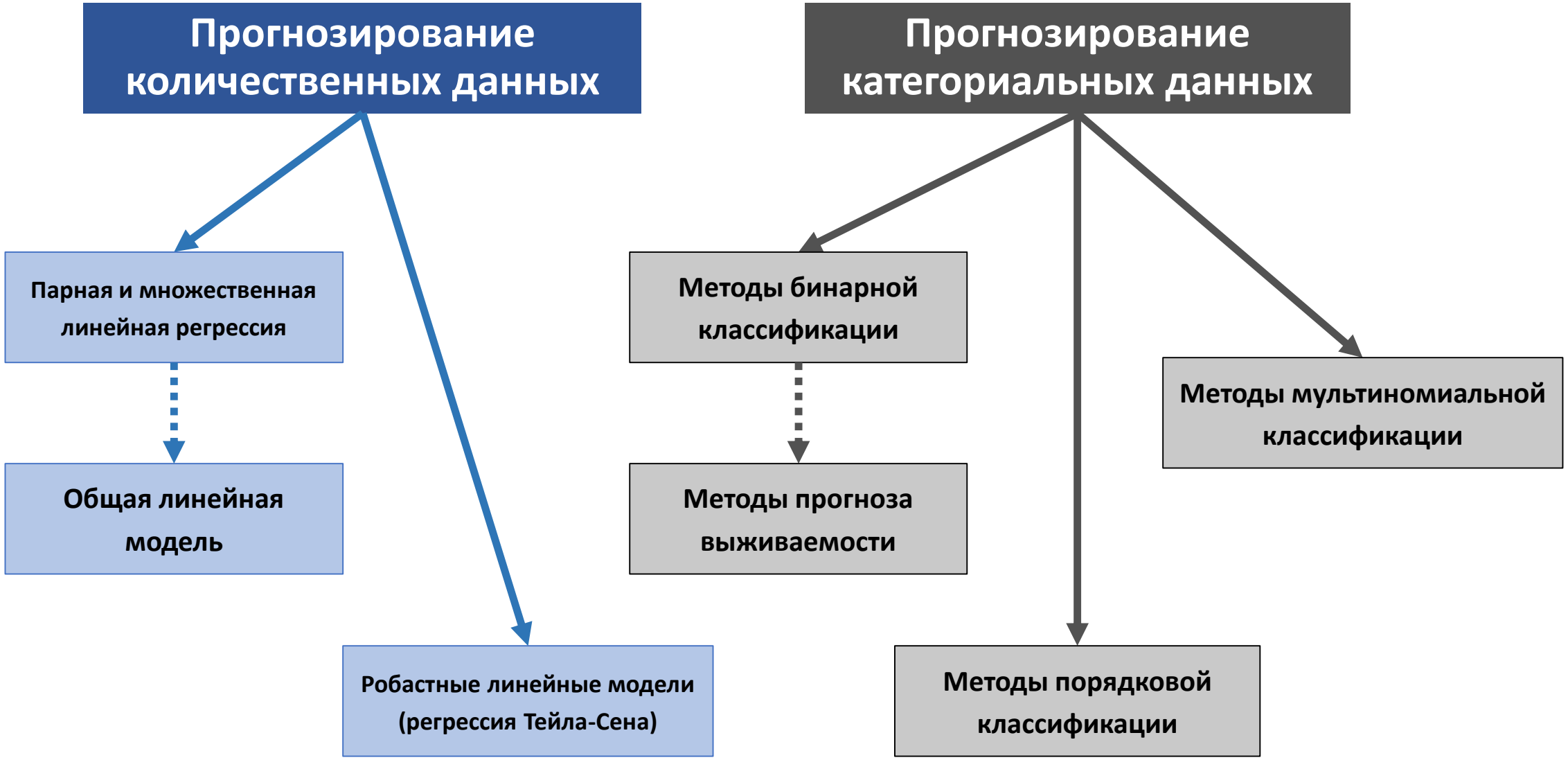
Использование прогностических моделей при анализе медицинских данных

МАРАПОВ ДАМИР ИЛЬДАРОВИЧ, к.м.н., доцент

@MEDSTATISTIC **MEDSTATISTIC.RU**

Виды прогностических моделей



Алгоритм выбора прогностической модели

УравнениЯ регрессии

Линейная регрессия
Общая линейная модель
Дискриминантный анализ

$$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + a_0$$

где y – зависимый показатель (исход, результат),
x – независимые показатели (предикторы),
$a_1 \ldots a_n$ – коэффициенты регрессии,
a_0 – интерсепт, величина y при $x_1, \ldots x_n = 0$

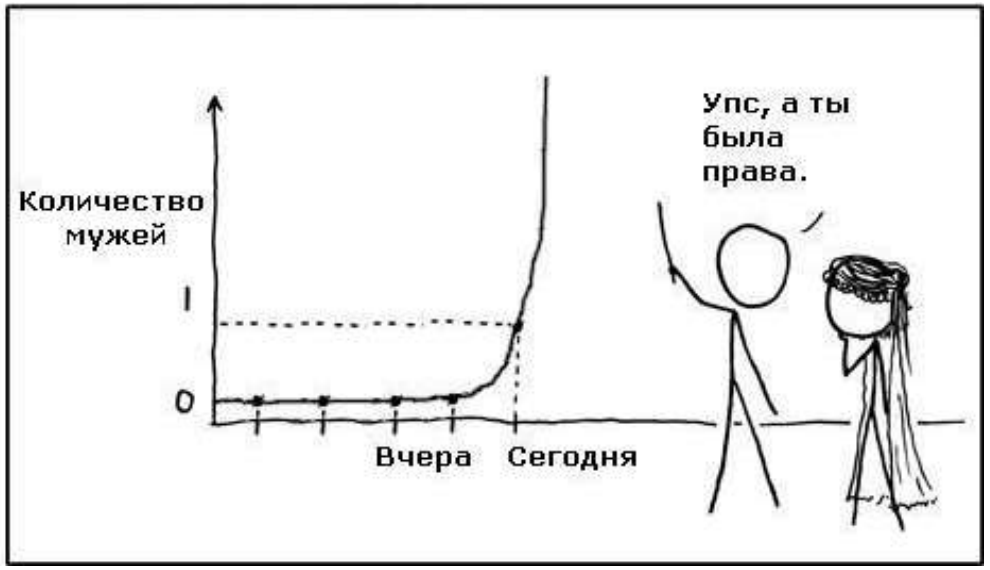
Логистическая регрессия
Регрессия Кокса
Порядковая регрессия

$$P = \frac{1}{1+e^{-(a_1x_1+a_2x_2+\ldots+a_nx_n+a_0)}}$$

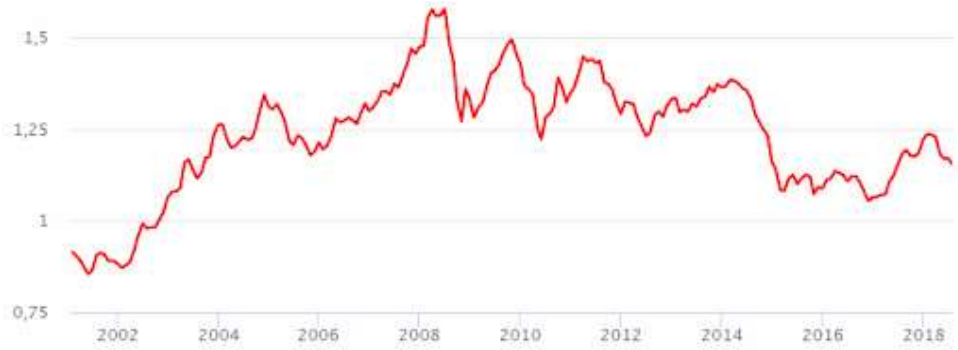
где: p – вероятность наступления исхода в долях единицы,
x – независимые показатели (предикторы),
$a_1 \ldots a_n$ – коэффициенты регрессии,
a_0 – интерсепт

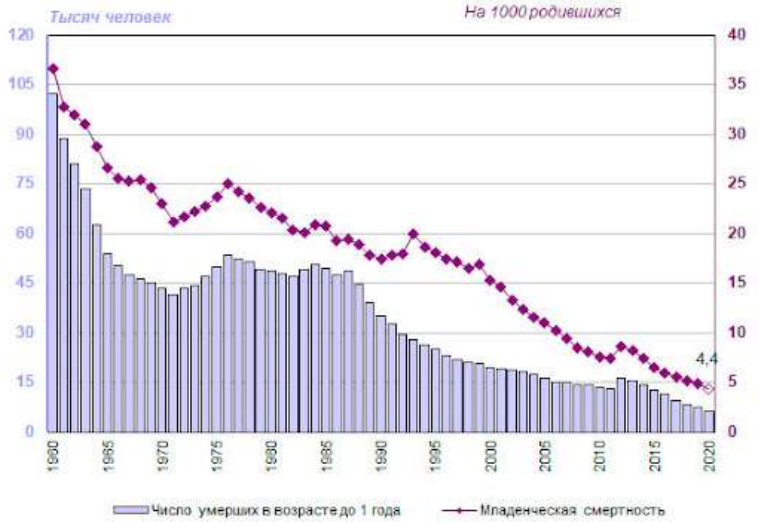
Медицинская статистика

Проблема экстраполяции в динамических рядах



EUR/USD





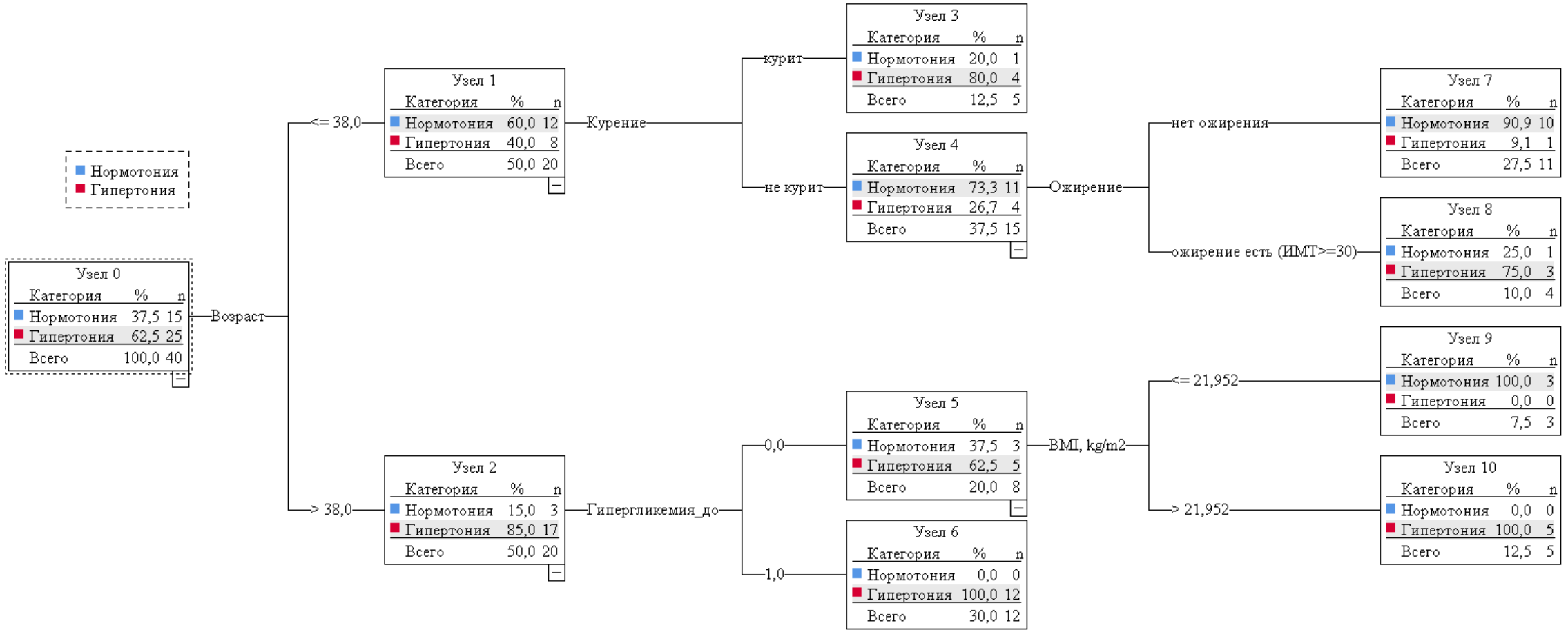
$Y = A\ (X_{max} + 1)$

$Y = UB\ 95\%\ CI\ (A) \cdot X_{max}$

Медицинская статистика

Деревья решений

Прогнозирование вероятности артериальной гипертонии в зависимости от наличия факторов риска



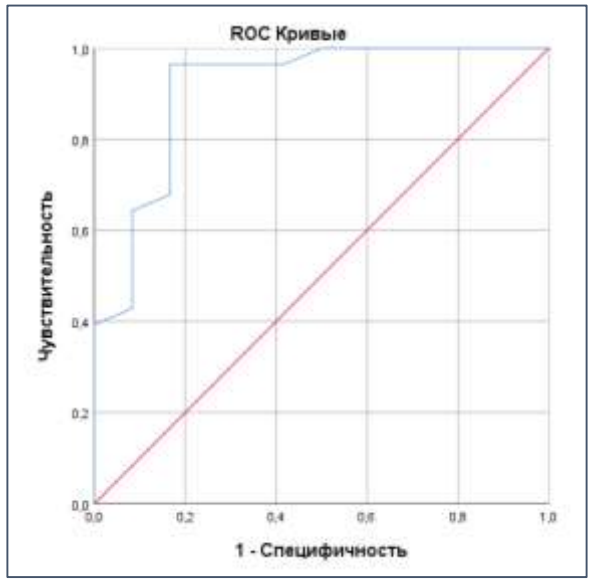
Медицинская статистика

Методы оценки прогностических моделей

- ✓ **Коэффициент корреляции (r_{xy} / ρ)** – характеризует направление и тесноту связи между ожидаемыми и наблюдаемыми значениями зависимой переменной
- ✓ **p-value** – уровень значимости изменения точности прогнозов при включении факторов в модель
- ✓ **R^2 (или псевдо-R^2)** – рассчитываемые различными способами коэффициенты детерминации, указывающие на долю дисперсии зависимой переменной, объясняемой полученной моделью
- ✓ **η^2 / ω^2** – частный вклад предиктора в дисперсию зависимой переменной
- ✓ **Чувствительность модели (Se)** – доля верно предсказанных случаев наличия исхода
- ✓ **Специфичность модели (Sp)** – доля верно предсказанных случаев отсутствия исхода
- ✓ **Отношение шансов (OR, odds ratio) или отношение рисков (HR, hazard ratio)** – характеризуют степень увеличения или уменьшения вероятности исхода при изменении значения предиктора

Медицинская статистика

Оценка прогностической значимости моделей бинарной классификации



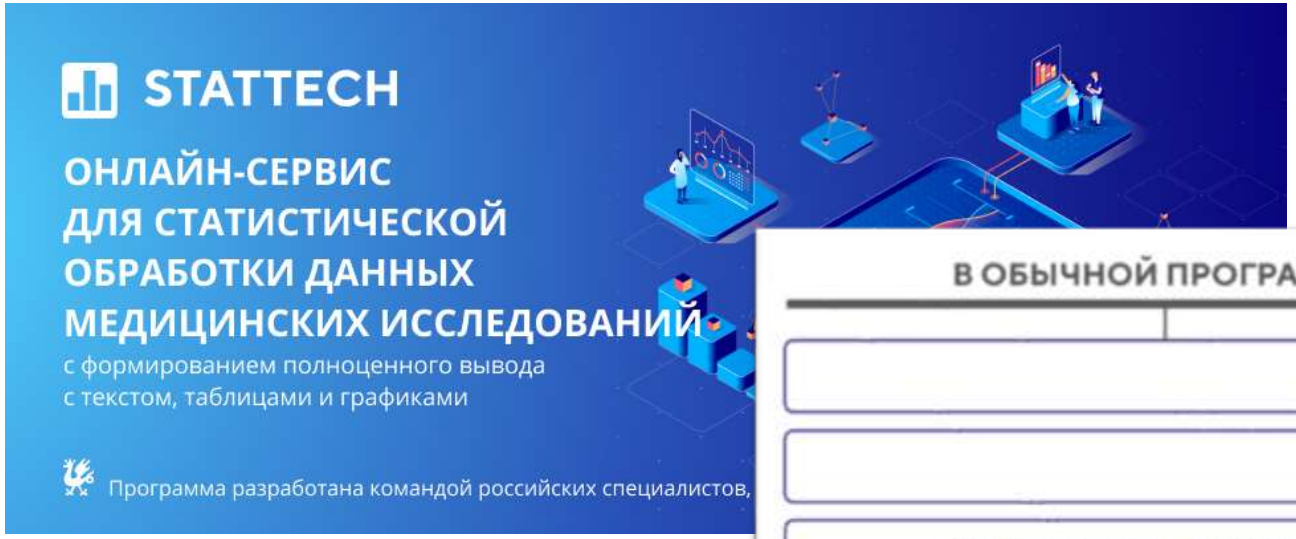
AUC – area under curve, площадь под кривой.
AUC ~ 0,5-1,0, LB 95% CI > 0,5

Фактический исход (из базы данных)	Прогнозируемый исход (высокий риск, исходя из предикторов)		
	АГ = 1? (Курение = 1)	АГ = 0? (Курение = 0)	
АГ = 1	**ИСТИННО +** **(true positive, TP)**	**ЛОЖНО –** **(false negative, FN)**	**Чувствительность** **Se = TP / (TP + FN)** «Истинно+» прогнозы / Все положительные исходы
АГ = 0	**ЛОЖНО +** **(false positive, FP)**	**ИСТИННО –** **(true negative TN)**	**Специфичность** **Sp = TN / (TN + FP)** «Истинно–» прогнозы / Все отрицательные исходы
	Положительная прогностическая значимость **PPV = TP / (TP + FP)** «Истинно+» прогнозы / Все положительные прогнозы	**Отрицательная прогностическая значимость** **NPV = TN / (TN + FN)** «Истинно–» прогнозы / Все отрицательные прогнозы	**Диагностическая эффективность** **(TP + TN) / N** Все истинные прогнозы / Общее число исследуемых

Медицинская статистика

В какой программе построить прогностические модели?

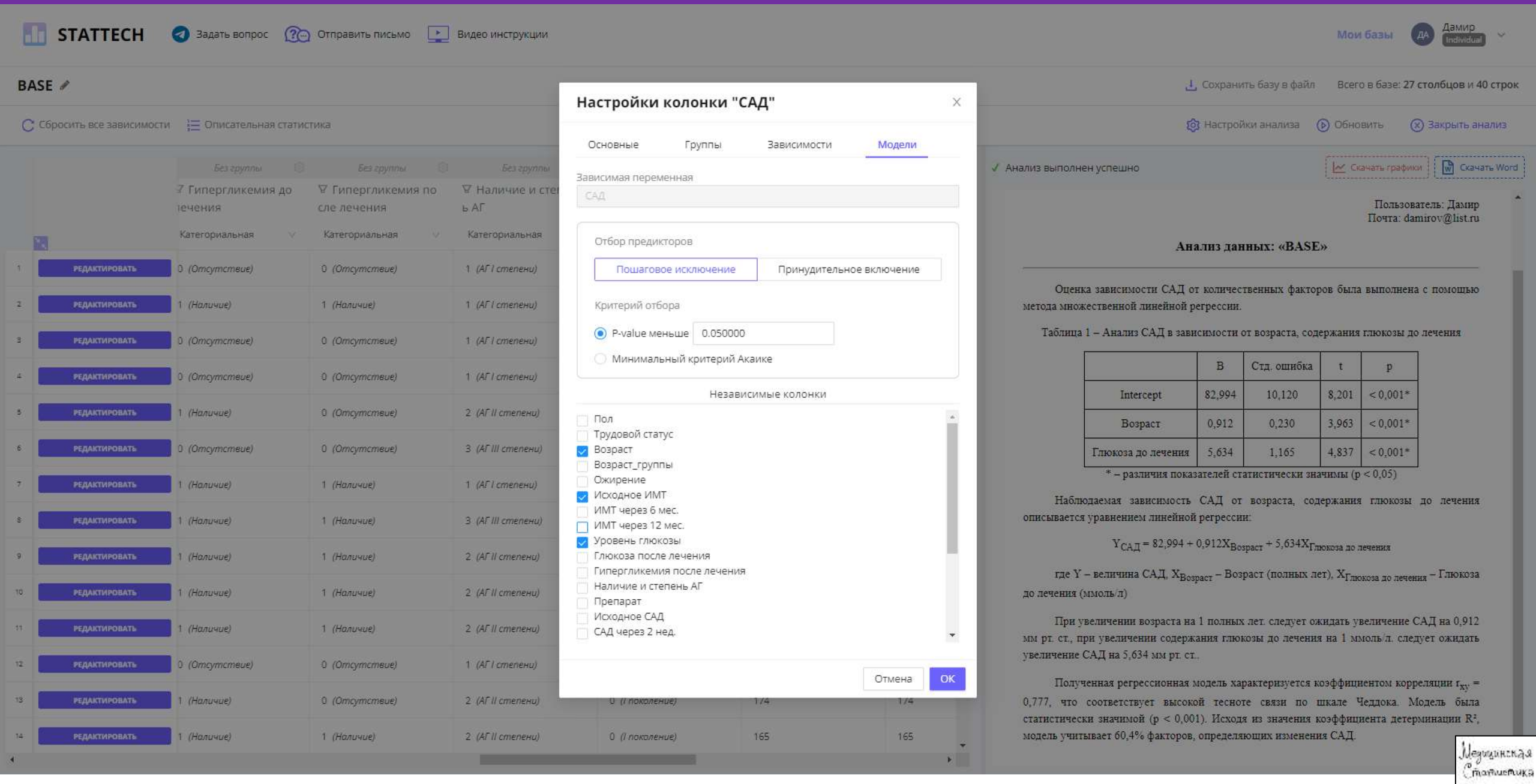
Построение регрессионных моделей в StatTech



В ОБЫЧНОЙ ПРОГРАММЕ	В ПРОГРАММЕ STATTECH
Подготовка базы данных в Excel	
Загрузка базы данных в программу	
Определение зависимых и независимых показателей исходя из задач исследования	
Выбор методов описания и анализа данных	нажать кнопку АНАЛИЗ
Получение набора числовых описательных и оценочных показателей	
Самостоятельная интерпретация	
Самостоятельное оформление таблиц, графиков и текста выводов	
Получение готового описания результатов исследования с таблицами, графиками и текстом	

Медицинская статистика

Настройка параметров модели в StatTech



Вывод линейной регрессии в StatTech

STATTECH | Задать вопрос | Отправить письмо | Видео инструкции | Мои базы | Дамир Individual

BASE | Сохранить базу в файл | Всего в базе: 27 столбцов и 40 строк

Сбросить все зависимости | Описательна | Настройки анализа | Обновить | Закрыть анализ

Был выполнен корреляционный анализ взаимосвязи возраста и САД.

Таблица 1 – Результаты корреляционного анализа взаимосвязи возраста и САД

Показатель	Характеристика корреляционной связи		
	r_{xy}	Теснота связи по шкале Чеддока	p
Возраст – САД	0,595	Заметная	< 0,001*

* – различия показателей статистически значимы (p ≤ 0,050)

Наблюдаемая зависимость САД от возраста описывается уравнением парной линейной регрессии:

$$Y_{САД} = 1,259 \times X_{Возраст} + 105,54$$

При увеличении возраста на 1 полных лет следует ожидать увеличение САД на 1,259 мм рт. ст. В соответствии с коэффициентом детерминации R^2 в полученной модели учтено 35,4% факторов, оказывающих влияние на значение САД.

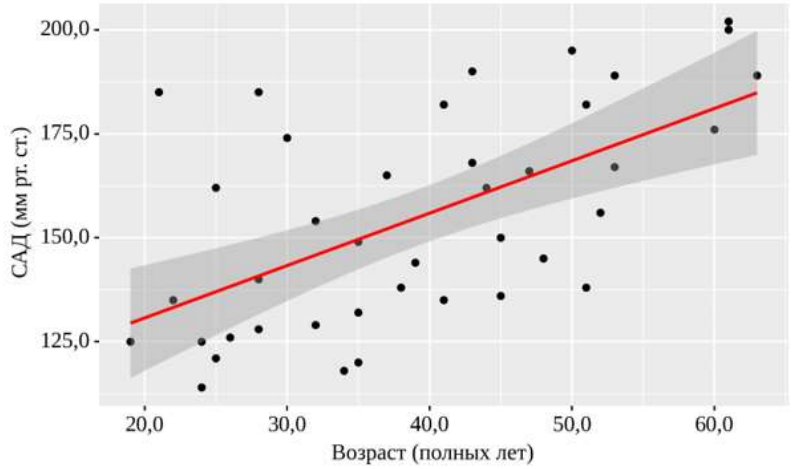


Рисунок 1 – График регрессионной функции, характеризующий зависимость САД от возраста

✓ Анализ выполнен успешно | Скачать графики | Скачать Word

Пользователь: Дамир
Почта: damirov@list.ru

Анализ данных: «BASE»

Оценка зависимости САД от количественных факторов была выполнена с помощью метода множественной линейной регрессии.

Таблица 1 – Анализ САД в зависимости от возраста, содержания глюкозы до лечения

	B	Стд. ошибка	t	p
Intercept	82,994	10,120	8,201	< 0,001*
Возраст	0,912	0,230	3,963	< 0,001*
Глюкоза до лечения	5,634	1,165	4,837	< 0,001*

* – различия показателей статистически значимы (p < 0,05)

Наблюдаемая зависимость САД от возраста, содержания глюкозы до лечения описывается уравнением линейной регрессии:

$$Y_{САД} = 82,994 + 0,912X_{Возраст} + 5,634X_{Глюкоза\ до\ лечения}$$

где Y – величина САД, $X_{Возраст}$ – Возраст (полных лет), $X_{Глюкоза\ до\ лечения}$ – Глюкоза до лечения (ммоль/л)

При увеличении возраста на 1 полных лет. следует ожидать увеличение САД на 0,912 мм рт. ст., при увеличении содержания глюкозы до лечения на 1 ммоль/л. следует ожидать увеличение САД на 5,634 мм рт. ст..

Полученная регрессионная модель характеризуется коэффициентом корреляции r_{xy} = 0,777, что соответствует высокой тесноте связи по шкале Чеддока. Модель была статистически значимой (p < 0,001). Исходя из значения коэффициента детерминации R^2, модель учитывает 60,4% факторов, определяющих изменения САД.

Медицинская статистика

Вывод бинарной логистической регрессии в StatTech

Была разработана прогностическая модель для определения вероятности наличия АГ в зависимости от наличия курения, содержания глюкозы в сыворотке крови методом бинарной логистической регрессии. Наблюдаемая зависимость описывается уравнением:

$P = 1 / (1 + e^{-z}) \times 100\%$

$z = -4{,}098 + 2{,}212X_{\text{Наличие курения}} + 0{,}629X_{\text{Уровень глюкозы}}$

где P – вероятность наличия АГ, $X_{\text{Наличие курения}}$ – Курение (0 – Отсутствие курения, 1 – Наличие курения), $X_{\text{Уровень глюкозы}}$ – Уровень глюкозы (ммоль/л)

Полученная регрессионная модель является статистически значимой ($p < 0{,}001$). Исходя из значения коэффициента детерминации Найджелкерка, модель учитывает 44,9% факторов, определяющих дисперсию вероятности наличия АГ.

Исходя из значений регрессионных коэффициентов, была установлена прямая связь содержания глюкозы в сыворотке крови с вероятностью выявления наличия АГ. Наличие курения при оценке влияния наличия курения сопровождалось увеличением вероятности наличия АГ.

Таблица 1 – Характеристики связи предикторов модели с вероятностью выявления наличия АГ

Предикторы	Unadjusted		Adjusted	
	COR; 95% ДИ	p	AOR; 95% ДИ	p
intercept	0,750	0,441	0,017	1,775
Курение: Наличие курения	7,111	0,768	9,131	0,875
Уровень глюкозы	1,701	0,239	1,876	0,291

* – влияние предиктора статистически значимо ($p < 0{,}05$)

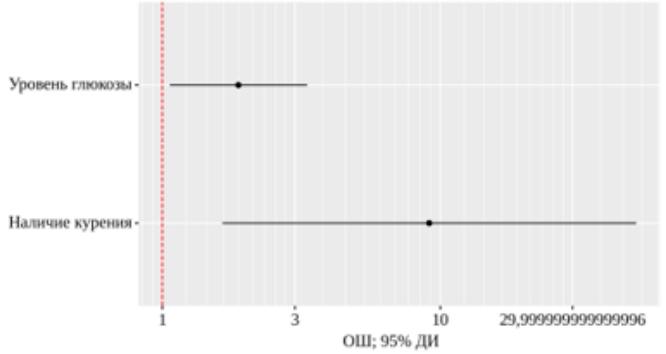


Рисунок 1 – Оценки отношения шансов с 95% ДИ для изучаемых предикторов наличия АГ

При оценке зависимости вероятности наличия АГ от значения логистической функции P с помощью ROC-анализа была получена следующая кривая.

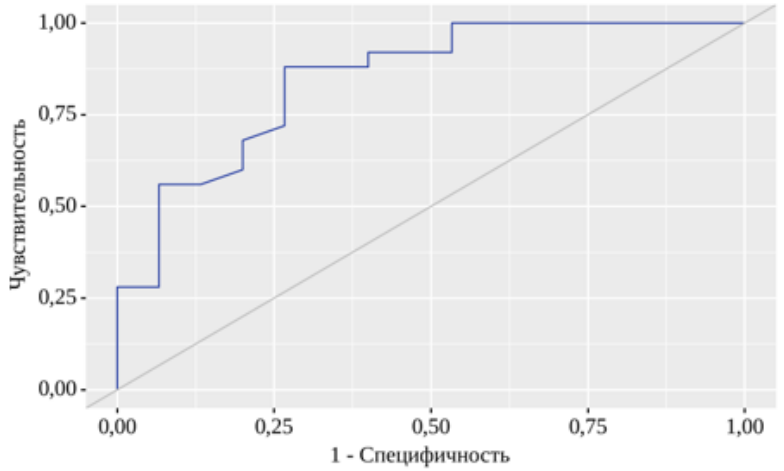


Рисунок 2 – ROC-кривая, характеризующая зависимость вероятности наличия АГ от значения логистической функции P

Площадь под ROC-кривой составила 0,848 ± 0,060 с 95% ДИ: 0,730 – 0,966. Полученная модель была статистически значимой ($p < 0{,}001$).

Пороговое значение логистической функции P в точке cut-off, которому соответствовало наивысшее значение индекса Юдена, составило 0,482. Наличие прогнозировалось при значении логистической функции P выше данной величины или равном ей. Чувствительность и специфичность модели составили 88,0% и 73,3%, соответственно.

Медицинская статистика

БЛАГОДАРЮ ЗА ВНИМАНИЕ!